Draft of an article to appear in *The MIT Encyclopedia of the Cognitive Sciences* (Rob Wilson and Frank Kiel, editors), Cambridge, Massachusetts: MIT Press, 1997.

Bounded Rationality

Jon Doyle
Massachusetts Institute of Technology
Laboratory for Computer Science
545 Technology Square
Cambridge, Massachusetts 02139
http://www.medg.lcs.mit.edu/doyle
doyle@mit.edu

January 20, 1998

BOUNDED RATIONALITY: rationality as exhibited by decision makers of limited abilities. The ideal of RATIONAL DECISION MAKING formalized in RATIONAL CHOICE THEORY, UTILITY THEORY, and the FOUNDATIONS OF PROBABILITY requires choosing so as to maximize a measure of expected utility that reflects a complete and consistent preference order and probability measure over all possible contingencies. This requirement appears too strong to permit accurate description of the behavior of realistic individual agents studied in economics, psychology, and artificial intelligence. Since rationality notions pervade approaches to so many other issues, finding more accurate theories of bounded rationality constitutes a central problem of these fields. Prospects appear poor for finding a single "right" theory of bounded rationality due to the many different ways of weakening the ideal requirements, some formal impossibility and tradeoff theorems, and the rich variety of psychological types observable in people, each with different strengths and limitations in reasoning abilities. The textbook of Russell and Norvig [17] provides a comprehensive survey of the roles of rationality and bounded rationality notions in artificial intelligence. Cherniak

[1] provides a philosophical introduction to the subject. Simon [21] discusses numerous topics in economics; see [2] for a broad economic survey.

Studies in ECONOMICS AND COGNITIVE SCIENCE and of human DECISION MAKING document cases in which everyday and expert decision makers do not live up to the rational ideal [11, 14]. The ideal maximization of expected utility implies a comprehensiveness at odds with observed failures to consider alternatives outside those suggested by the current situation. The ideal probability and utility distributions imply a degree of LOGICAL OMNISCIENCE that conflicts with observed inconsistencies in beliefs and valuations and with the frequent need to invent rationalizations and preferences to cover formerly unconceived circumstances. The theory of BAYESIAN LEARNING or conditionalization, commonly taken as the theory of belief change or learning appropriate to rational agents, conflicts with observed difficulties in assimilating new information, especially the resistance to changing cognitive habits.

Reconciling the ideal theory with views of decision makers as performing computations also poses problems. Conducting the required optimizations at human rates using standard computational mechanisms, or indeed any physical system, seems impossible to some. The seemingly enormous information content of the required probability and utility distributions may make computational representations infeasible, even using BAYESIAN NETWORKS or other relatively efficient representations.

The search for realistic theories of rational behavior began by relaxing optimality requirements. Simon [19] formulated the theory of "satisficing", in which decision makers seek only to find alternatives that are satisfactory in the sense of meeting some threshold or "aspiration level" of utility. A more general exploration of the idea of meeting specific conditions rather than unbounded optimizations also stimulated work on PROBLEM SOLVING, which replaces expected utility maximization with acting to satisfy sets of goals, each of which may be achieved or not. Simon [20] also emphasized the distinction between "substantive" and "procedural" rationality, concerning respectively rationality of the result and of the process by which the result was obtained, setting procedural rationality as a more feasible aim than substantive rationality. Good [8, 9] urged a related distinction in which "Type 1" rationality consists of the ordinary ideal notion, and "Type 2" rationality consists of making ideal decisions taking into account the cost of deliberation. The Simon and Good distinctions informed work in artificial

intelligence on control of reasoning [3], including explicit deliberation about the conduct of reasoning [5], economic decisions about reasoning [10, 16], and iterative approximation schemes or "anytime algorithms" [10, 4] in which optimization attempts are repeated with increasing amounts of time, so as to provide an informed estimate of the optimal choice no matter when deliberation is terminated. Although reasoning about the course of reasoning may appear problematic, it may be organized to avoid crippling circularities (see METAREASONING), and admits theoretical reductions to nonreflective reasoning [13]. One may also relax optimality by adjusting the scope of optimization as well as the process. Savage [18] observed the practical need to formulate decisions in terms of "small worlds" abstracting the key elements, thus removing the most detailed alternatives from optimizations. The related notions of "selective rationality" [12] and "bounded optimality" [10, 15] treat limitations stemmings from optimization over circumscribed sets of alternatives.

Lessening informational requirements constitutes one important form of procedural rationality. Goal-directed problem solving and small world formulations do this directly by basing actions on highly incomplete preferences and probabilities. The extreme incompleteness of information represented by these approaches can prevent effective action, however, thus requiring means for filling in critical gaps in reasonable ways, including various JUDGMENT HEURISTICS based on representativeness or other factors [11]. Assessing the expected value of information forms one general approach to filling these gaps. In this approach, one estimates the change in utility of the decision that would stem from filling specific information gaps, and then acts to fill the gaps offering the largest expected gains. These assessments may be made of policies as well as of specific actions. Applied to policies about how to reason, such assessments form a basis for the nonmonotonic or default reasoning methods appearing in virtually all practical inference systems (formalized as various NONMONOTONIC LOGICS and theories of belief revision) that fill routine gaps in rational and plausible ways. Even when expected deliberative utility motivates use of a nonmonotonic rule for adopting or abandoning assumptions, such rules typically do not involve probabilistic or preferential information directly, though some rules admit natural interpretations as either statements of extremely high probability (infinitesimally close to 1), in effect licensing reasoning about magnitudes of probabilities without requiring quantitative comparisons, or as expressions of preferences over beliefs and

other mental states of the agent, in effect treating reasoning as seeking mental states that are Pareto optimal with respect to the rules [6]. Nonmonotonic reasoning methods also augment BAYESIAN LEARNING (conditionalization) with direct changes of mind that suggest "conservative" approaches to reasoning that work through incremental adaptation to small changes, an approach seemingly more suited to exhibiting procedural rationality than the full and direct incorporation of new information called for by standard conditionalization.

Formal analogs of Arrow's impossibility theorem for social choice problems and multiattribute UTILITY THEORY limit the procedural rationality of approaches based on piecemeal representations of probability and preference information [7]. As such representations dominate practicable approaches, one expects any automatic method for handling inconsistencies amidst the probability and preference information to misbehave in some situations.

References

[1] C. Cherniak. *Minimal Rationality*. MIT Press, Cambridge, 1986.

[2] J. Conlisk. Why bounded rationality? *Journal of Economic Literature*, XXXIV:669–700, June 1996.

[3] T. Dean. Decision-theoretic control of inference for time-critical applications. *International Journal of Intelligent Systems*, 6(4):417–441, 1991.

[4] T. Dean and M. Boddy. An analysis of time-dependent planning. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 49–54, 1988.

[5] J. Doyle. A model for deliberation, action, and introspection. AI-TR 581, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA, 02139, 1980.

[6] J. Doyle. Reasoned assumptions and rational psychology. *Fundamenta Informaticae*, 20(1-3):35–73, 1994.

[7] J. Doyle and M. P. Wellman. Impediments to universal preference-based default theories. *Artificial Intelligence*, 49(1-3):97–128, May 1991.

[8] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society B*, 14:107–114, 1952.

[9] I. J. Good. The probabilistic explication of information, evidence, surprise, causality, explanation, and utility. In V. P. Godambe and D. A. Sprott, editors, *Foundations of Statistical Inference*, pages 108–127. Holt, Rinehart and Winston, Toronto, 1971.

[10] E. J. Horvitz. Reasoning about beliefs and actions under computational resource constraints. In *Proceedings of the Third AAAI Workshop on Uncertainty in Artificial Intelligence*. AAAI, 1987.

[11] D. Kahneman, P. Slovic, and A. Tversky, editors. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982.

[12] H. Leibenstein. *Beyond Economic Man: A New Foundation for Microeconomics*. Harvard University Press, Cambridge, MA, second edition, 1980.

[13] B. L. Lipman. How to decide how to decide how to . . .: modeling limited rationality. *Econometrica*, 59(4):1105–1125, 1991.

[14] M. J. Machina. Choice under uncertainty: Problems solved and unsolved. *Journal of Economic Perspectives*, 1(1):121–154, Summer 1987.

[15] S. Russell and D. Subramanian. Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2:575–609, 1995.

[16] S. J. Russell. *Do the Right Thing: Studies in Limited Rationality*. MIT Press, Cambridge, MA, 1991.

[17] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 1995.

[18] L. J. Savage. *The Foundations of Statistics*. Dover Publications, New York, second edition, 1972.

[19] H. A. Simon. A behavioral model of rational choice. *Quarterly Journal of Economics*, 69:99–118, 1955.

[20] H. A. Simon. From substantive to procedural rationality. In S. J. Latsis, editor, *Method and Appraisal in Economics*, pages 129–148. Cambridge University Press, 1976.

[21] H. A. Simon. *Models of Bounded Rationality: Behavioral Economics and Business Organization*, volume 2. MIT Press, Cambridge, MA, 1982.